Genetics and population analysis

REJECTOR: software for population history inference from genetic data via a rejection algorithm

Matthew J. Jobin1,∗ and Joanna L. Mountain1,2

1Department of Anthropology, Stanford University, Stanford, CA 94305 and 223andMe, Inc., Mountain View, CA, USA

Received on July 4, 2008; revised on September 16, 2008; accepted on October 15, 2008

1 INTRODUCTION

The rejection algorithm method can be an efficient way to infer distributions of genetic and demographic parameters using polymorphism data (Ramakrishnan et al., 2004). In this approach values for a parameter of interest are simulated from a prior distribution and accepted with a probability proportional to their likelihood. Since likelihoods are difficult to compute directly, approximate simulation methods have been developed wherein summary statistics are used in place of the full dataset, and the candidate parameter value is accepted based on the values of the statistics (Tavaré et al., 1997). Bayesian statistics incorporate the use of prior information in the form of distributions of parameters of interest. The rejection algorithm method has been extended by simulation of parameter values from prior distributions and by coalescent simulation of gene genealogies given these parameter values (Pritchard et al., 1999). Summary statistics calculated from these simulated genealogies can be compared with summary statistics calculated from observed data, and if the simulated summary statistics fall within a specified tolerance of the observed summary statistics, the parameter values are accepted to form a sample from the posterior distribution. The combination of summary statistics in place of the full dataset and Bayesian prior distributions is termed approximate Bayesian computation.

Approximate Bayesian methods using summary statistics can deal with complex models and provide estimates of parameter values of interest under any model (Excoffier and Heckel, 2006). REJECTOR implements an approximate Bayesian method termed rejection-based approximate Bayesian inference (Beaumont et al., 2002), for inferring population history.

REJECTOR calculates summary statistic values from observed data, then simulates a series of population histories from prior distributions for parameters of interest (e.g. time of divergence, population size and migration rate) and calculates summary statistic values for each simulated population history. REJECTOR accepts a simulated population history if its summary statistic values fall within a user-specified tolerance value of the values calculated from the experimental data. The parameter values of the accepted histories form posterior distributions that can be used to make estimates of the true parameter values. In addition to parameter estimation, REJECTOR can be used to compare alternate models of population history by comparing the proportion of accepted simulated histories generated by each model.

REJECTOR uses any number of unlinked blocks of genetic data, each containing any number and combination of SNP/UEP, microsatellite and sequence loci. REJECTOR makes use of the extant SIMCOAL2 (Laval and Excoffier, 2004) software for coalescent simulation, an updated version that allows for multiple blocks of linked loci. Comparisons can be made using a wide array of summary statistics. REJECTOR allows the user to jointly analyze different categories of genetic data simultaneously. Because each iteration of a rejection algorithm-based simulation is independent of all others, the computational work can be divided amongst multiple processors. A recent advance in Bayesian simulation, sequential Monte Carlo promises improved efficiency in rejection-based methods via the propagation of parameter values through a series of intermediate distributions with decreasing tolerance values (Sisson et al., 2007), but this iteration of the REJECTOR software does not include this feature.

2 METHODS

REJECTOR is a package composed of four executables, one of which is the SIMCOAL2 software by Laval and Excoffier. We have written the other three executables—REJSTATS, PRIOR and REJECTOR—in C++. We have compiled, tested and run the software under Mac OS X, Linux and Windows.

We tested the REJECTOR software package using simulated data in order to evaluate its ability to recover known parameter values using various configurations of data. This method uses summary statistics calculated from
As demonstrated above, the REJECTOR software can provide a simple two-population model with no growth, with equal population sizes of 1000 chromosomes and a time of divergence of 2000 generations. We used these summary statistic values as a target for a wide array of models using various data types, numbers of markers, tolerance levels and summary statistics. For each of these recovery tests, we ran REJECTOR for 100,000 iterations and compared the resultant statistics to those calculated from the target model, providing information on the effect that any combination of these factors has on the precision with which REJECTOR can recover a known set of parameter values. We have recorded the results of these recovery tests in the Supplementary Material, providing the researcher with guidelines for choosing combinations of summary statistics and data types. This information can serve as a starting point for the design of simulations that evaluate models given the experimentally derived data.

3 RESULTS

An example of the recovery test method described above demonstrates the precision and accuracy of REJECTOR using well-chosen summary statistics and data types. We simulated 100 unlinked microsatellites under the target model described in Section 2 (two populations of 1000 chromosomes each, no growth, time of divergence of 2000 generations and a mutation rate of $1.5 \times 10^{-3}$ per microsatellite per generation) and recorded the values for the summary statistics $\delta \mu^2$, the average squared distance in microsatellite lengths (Goldstein et al., 1995) and the average variance in microsatellite repeat lengths, or microsatellite variance. We then ran REJECTOR for 100,000 iterations under a model introducing Gaussian prior distributions for time of divergence and size of Population 1 whose means were not equal to the true parameter values. We accepted iterations into the posterior distribution if both $\delta \mu^2$ and microsatellite variance matched the target values within a tolerance of 0.05.

Posterior distributions can be presented in 2D density plots to show the response of each summary statistic to changes in any two parameters. Figure 1 displays the posterior distribution from the recovery test described above. This plot shows the posterior distribution centered on the correct time of divergence and size of Population 1. As recorded in the Supplementary Material, $\delta \mu^2$ is among the most sensitive summary statistics in the REJECTOR package to time of divergence, while microsatellite variance is sensitive to population size. The combination of these two summary statistics allows REJECTOR to recover the correct parameter values with the precision shown.

4 DISCUSSION

As demonstrated above, the REJECTOR software can provide estimates of parameter values for multiple parameters simultaneously, with both precision and accuracy if optimal summary statistics and tolerance values are chosen. Given sufficient computation time and data, the software has the ability to distinguish between alternate models of population history through both posterior distributions and the comparison of accepted runs for each parameter and each model (Supplementary Material). Hypotheses of ancient human migration can, for example, be investigated using REJECTOR by creating two alternate models with very different migration rates (e.g. one model with a hypothesized rate and one with no migration at all) and comparing acceptance rate (Pritchard et al., 1999).

Fig. 1. Density plot of time of divergence versus size of Population 1 for 100 unlinked short tandem repeats (STRs) using the summary statistics $\delta \mu^2$ and microsatellite variance. The light gray contours represent prior distributions for time of divergence (mean = 2250 generations, SD = 500) and size of Population 1 (mean = 1250 chromosomes, SD = 500). The darker contours represent the posterior distribution of parameter values collected from 100,000 iterations of REJECTOR with a tolerance of 0.05. The small black circle marks the true values of the parameters—2000 generations for time of divergence and 1000 chromosomes for size of Population 1.

ACKNOWLEDGEMENTS

We thank L. Excoffier for his comments on the draft version of this article.

Funding: Morrison Institute for Population Resource Studies (to M.J.J.); National Institutes of Health (GM028428 to J.L.M.).

Conflict of Interest: none declared.

REFERENCES


Tavaré et al. (1997) Inferring coalescence times from DNA sequence data. Genetics, 145, 505–518.